CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

A Nested Recursive Logit Model for Route Choice Analysis

Anh Tien Mai
Mogens Fosgerau
Emma Frejinger

August 2014

CIRRELT-2014-39

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

A Nested Recursive Logit Model for Route Choice Analysis

Anh Tien Mai[1,*], Mogens Fosgerau[2], Emma Frejinger[1]

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT) and Department of Computer Science and Operations Research, Université de Montréal, P.O. Box 6128, Station Centre-Ville, Montréal, Canada H3C 3J7

[2] Technical University of Denmark, Bygningstorvet, Building 116B, room 123A, 2800 Kgs. Lyngby, Denmark, and Royal Institute of Technology, SE-100 44, Stockholm, Sweden

Abstract. Recently, Fosgerau et al. (2013) proposed a route choice model called recursive logit (RL) model that can be consistently estimated based on revealed preference data and used for prediction without sampling any choice sets of paths. They show that the RL model is equivalent to a logit model and hence it exhibits the independence from irrelevant alternatives (IIA) property although it is reasonable to assume that it does not hold in a route choice context. This paper presents the nested recursive logit (NRL) model that relaxes the IIA assumption by allowing scale parameters to be link specific while keeping the advantages of the RL model. The key challenge lies in the computation of the expected maximum utility from a position in the network to a destination (value functions). In the RL model the value functions can be efficiently computed by solving a system of linear equations. In the case of NRL they are the solution to a system of non-linear equations which is considerably more difficult to deal with. We propose an iterative method with dynamic accuracy that makes it possible to estimate and apply the NRL efficiently in real networks. We report estimation results and a prediction study for a network composed of more than 3000 nodes and 7000 links. The results show that the NRL model has sensible parameter estimates and the bit is remarkably better than the RL model. They are more similar in the prediction performance.

Keywords: Route choice model, recursive logit, nested recursive logit, substitution patterns, value iterations, maximum likelihood estimation.

Acknowledgements. This research was partially funded by Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant 435678-2013.

---

* Corresponding author: AnhTien.Mai@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2014

1 Introduction

Discrete choice models are generally used for analyzing path choices in real networks based on revealed preference (RP) data. There are two main modelling issues associated with consistently estimating such models and subsequently using them for prediction. First, the choice sets are unknown to the analyst and the set of all feasible paths for a given origin-destination pair cannot be enumerated. Second, path utilities may be correlated, for instance, due to physical overlap in the network. There is no state-of-the-art path choice model that can be both consistently estimated and used for prediction while addressing the two aforementioned issues. The objective of this paper is to fill this gap in the literature by proposing a new model that we call nested recursive logit.

Most of the existing path choice models are based on choice sets of paths that need to be sampled before estimating or applying the model. Many different algorithms exist for sampling choice sets (for reviews, see e.g. Frejinger et al., 2009, Prato, 2009) and they all correspond to importance sampling protocols where paths have non-equal probabilities of being sampled. Frejinger et al. (2009) argue that utilities need to be corrected for the sampling of alternatives which implies that only algorithms that allow to compute path sampling probabilities can be used. Frejinger et al. (2009) use the logit (MNL) model but other types of generalized extreme value (GEV) models can be used (Guevara and Ben-Akiva, 2013) although it has not yet been done for path choice. It is important to note that the sampling approach can be used to consistently estimate a path choice model but it is still unknown how to use the model for prediction.

A number of models in the literature allow to model the correlation structure of path utilities. A few examples are the link-nested logit (Vovsha and Bekhor, 1998), mixed logit with error components (Bekhor et al., 2001, Frejinger and Bierlaire, 2007) and paired combinatorial logit (Chu, 1989). These models are based on sampled choice sets without correcting the utilities for the sampling protocol. Hence, the parameter estimates are conditional on the choice sets and may have significantly different values if some paths are added or removed from the choice sets. This is problematic since the true choice sets are unknown. As mentioned earlier, the GEV models (e.g. link-nested logit) can be corrected, while it is unclear how to do it for the mixed logit models.

Recently, Fosgerau et al. (2013) proposed the recursive logit (RL) model where path choice is modelled as a sequence of link choices using a dynamic discrete choice framework. The RL model can be consistently estimated and used for prediction without sampling choice sets of paths. It is however

equivalent to a MNL model over the set of all feasible paths and even though a correction attribute called link size, similar to path size, is proposed, it cannot properly model correlated path utilities.

In this paper we propose an extension of the RL model that allows to model correlated path utilities in a fashion similar to nested logit (Ben-Akiva, 1973) where links can have different scale parameters. The key challenge lie in the computation of the expected maximum utility from a current position in the network until the destination (value functions). The strength of the RL model is that the value functions can be computed by solving a system of linear equations, which is fast and easy to do. In the case of the nested RL (NRL), the value functions are a solution to a system of *non-linear* equations which is substantially more difficult to deal with. We propose an iterative method with dynamic accuracy to efficiently solve this system.

This paper makes a number of contributions. First we propose a model that can be consistently estimated and used for prediction without sampling choice sets while allowing to model correlated path utilities. Second, we provide illustrative examples and discuss the resulting substitution patters. Third, we propose an iterative method to solve the value functions and derive the analytical gradient of the log-likelihood function for the case that the scales are functions of model parameters so that the NRL model can be efficiently estimated. Fourth, we present estimation results based on real data for a network with 3000 nodes and 7000 links. Finally, the estimation code is implemented in MATLAB and is freely available upon request.

The paper is structured as follows. Section 2 presents the NRL model. Section 3 discusses substitution patterns by illustrative examples and Section 4 provide a method to compute the value functions. Section 5 derives an analytical formula for the first order derivative of the log-likelihood function. Specifications, estimation and prediction results are presented in Section 6 and finally Section 7 concludes.

2 The nested recursive logit model

The RL model is recently proposed by Fosgerau et al. (2013) where the path choice problem is formulated as a sequence of link choices and modeled in a dynamic discrete choice framework. They consider the case where the random terms are independently and identically distributed (i.i.d.) extreme value type I with zero mean so that the model is equivalent to MNL. In this section we present the NRL model which relaxes the independence of irrelevant alternatives (IIA) property of MNL by assuming that the scales of random terms are different over links. In the following we derive the model

using the same notation as Fosgerau et al. (2013). Even though the derivation is similar to the RL case, the resulting path and link choice probabilities are different.

A directed connected graph (not assumed acyclic) $G = (A; \mathcal{V})$ is considered, where $A, \mathcal{V}$ are the set of links and nodes, respectively. For each link $k \in A$, we denote the set of outgoing links from the sink node of k by $A(k)$. Moreover we associate an absorbing state with each destination by extending the network with dummy links (see Figure 1). The set of all links is $\tilde{A} = A \cup \{d\}$ and the corresponding deterministic utility is $v(d|k) = 0$ for all k that have destination d as sink node. Given two links $a, k \in \tilde{A}$, the following instantaneous utility is associated with action $a \in A(k)$

$$u(a|k;\beta) = v(a|k;\beta) + \mu_k \epsilon(a)$$

where β is a vector of parameters, μ_k is a strictly positive scale parameter and $\epsilon(a)$ are i.i.d extreme value type I with zero mean. The deterministic term $v(a|k;\beta)$ is assumed negative for all links except the dummy link d. We emphasize the difference with the original RL model where scale parameters are assumed equal ($\mu_k = \mu \; \forall k \in A$).

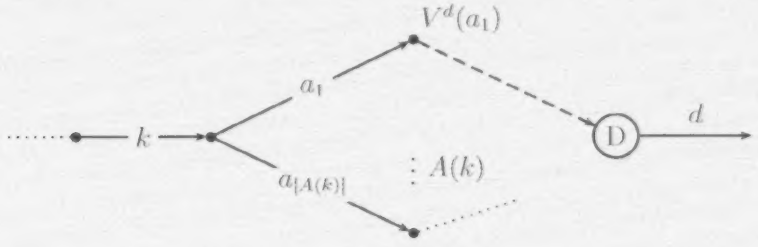

Figure 1: Illustration of notation (Fosgerau et al., 2013)

The expected maximum utility from the sink node of a link k to the destination is the value function $V^d(k;\beta)$. The superscript d indicates that the value functions are destination specific and they also depend on parameters β. $V^d(k;\beta)$ is recursively defined by Bellman's equation

$$V^d(k;\beta) = \mathbb{E}\Big[\max_{a \in A(k)} \big(v(a|k;\beta) + V^d(a;\beta) + \mu_k \epsilon(a)\big)\Big] \quad \forall k \in A \qquad (1)$$

or equivalently

$$\frac{1}{\mu_k} V^d(k;\beta) = \mathbb{E}\Big[\max_{a \in A(k)} \big(\frac{1}{\mu_k}(v(a|k;\beta) + V^d(a;\beta)) + \epsilon(a)\big)\Big] \quad \forall k \in A. \qquad (2)$$

For notational simplicity we omit from now on β from the value functions $V(.)$ and the utilities $v(.)$. For the same reason, we omit an index for individual n but note that utilities $u(a|k)$ can be individual specific.

Given these assumptions the probability of choosing a given state k is given by the MNL model

$$\begin{aligned} P^d(a|k) &= \delta(a|k)\frac{e^{\frac{1}{\mu_k}(v(a|k)+V^d(a))}}{\sum_{a'\in A(k)} e^{\frac{1}{\mu_k}(v(a'|k)+V^d(a'))}} \\ &= \delta(a|k)e^{\frac{1}{\mu_k}(v(a|k)+V^d(a)-V^d(k))} \quad \forall k, a \in \tilde{A} \end{aligned} \tag{3}$$

where $\delta(a|k) = 1$ if $a \in A(k)$ and zero otherwise. The value functions in this case are given by the logsum

$$\frac{1}{\mu_k}V^d(k) = \ln\Big(\sum_{a\in A(k)} e^{\frac{1}{\mu_k}(v(a|k)+V^d(a))}\Big) \quad \forall k \in A \tag{4}$$

and $V^d(d) = 0$ by assumption. Similar to Fosgerau et al. (2013) we can write Equation (4) as

$$e^{\frac{1}{\mu_k}V^d(k)} = \begin{cases} \sum_{a\in A}\delta(a|k)e^{\frac{(v(a|k)+V^d(a))}{\mu_k}} & \forall k \in A \\ 1 & k = d \end{cases} \tag{5}$$

and define a matrix $M^d(|\tilde{A}| \times |\tilde{A}|)$ and a vector $z^d(|\tilde{A}| \times 1)$ with entries

$$M^d_{ka} = \delta(a|k)e^{\frac{v(a|k)}{\mu_k}}, \; z^d_k = e^{\frac{V(k)}{\mu_k}}, \quad k, a \in \tilde{A}. \tag{6}$$

The key issue here compared to the RL model is that we do not end up with a system of linear equations. Indeed, the value functions are the solutions to the following system of non-linear equations

$$z^d_k = \begin{cases} \sum_{a\in A} M^d_{ka}(z^d_a)^{\mu_a/\mu_k} & \forall k \in A \\ 1 & k = d. \end{cases} \tag{7}$$

Moreover the probability of a path σ defined by a sequence of links $\sigma = [k_0, k_1, \ldots, k_I]$ has also a slightly more complicated expression than the RL path probability because link specific value functions do not cancel due to the scale parameters

$$P(\sigma) = \prod_{i=0}^{I-1} e^{\frac{1}{\mu_{k_i}}(v(k_{i+1}|k_i)+V^d(k_{i+1})-V^d(k_i))}. \tag{8}$$

Finally we note that the IIA property does not hold. Consider the ratio of the choice probabilities of two paths $\sigma_1 = [k_1, \ldots, k_{I_1}]$ and $\sigma_2 = [h_1, \ldots, h_{I_2}]$ connecting a same origin-destination pair

$$\frac{P(\sigma_1)}{P(\sigma_2)} = \frac{\prod_{i=1}^{I_1-1} e^{\frac{1}{\mu_{k_i}}(v(k_{i+1}|k_i)+V^d(k_{i+1})-V^d(k_i))}}{\prod_{i=1}^{I_2-1} e^{\frac{1}{\mu_{h_i}}(v(h_{i+1}|h_i)+V^d(h_{i+1})-V^d(h_i))}}. \tag{9}$$

When the scales $\mu_k = \mu \ \forall k \in A$, the value functions cancel out and the ratio (9) only depends on the utilities of two considered paths. For the NRL model, the ratio (9) depends on several values functions, which are evaluated based on the whole network and therefore the IIA property does not hold. In the following section be discuss the resulting substitution pattern in more depth using two illustrative examples.

3 Illustrative examples and substitution patterns

First we consider the small network shown in Figure 2 which is designed so that each path in the network belong to exactly one nest when defined by physical overlap (the nesting structure is shown in the figure). There are 4 nodes A, B, C, D and 9 links (link o is the origin and link d is a dummy link). Moreover, there are 6 possible paths from o to d: $[o, a, a_1, d]$, $[o, a, a_2, d]$, $[o, a, a_3, d]$, $[o, b, b_1, d]$, $[o, b, b_2, d]$ and $[o, b, b_3, d]$ and we number these paths as 1, 2, 3, 4, 5 and 6, respectively. The only attribute in the instantaneous utility is link length and the values are given in the parentheses on each arc. In order to compute path probabilities we choose a length parameter $\tilde{\beta} = -1$.

For the RL model the IIA property holds meaning that if we remove links in the network, the probabilities of feasible paths will increase by the same proportions (for example if we remove link a_2, the probabilities of path $[o, a, a_3, d]$ and path $[o, b, b_3, d]$ increase but they are still equal). For the NRL model, the scales of random terms are assigned different values. To evaluate the impact of the scales on the path probabilities we assign a scale of 0.5 for links a, a scale of 0.8 for links b and a scale of 1.0 for the others. Similar to an example in Train (2003), we illustrate substitution patterns by removing in turn links a_1, a_2, b_1, b_2 and present changes in probabilities in Table 1.

We note that the changes in probabilities for paths $[o, a, a_1, d]$, $[o, a, a_2, d]$, $[o, a, a_3, d]$ rise by the same proportions whenever one link is removed from the network. This is also the case for the three paths $[o, b, b_1, d]$, $[o, b, b_2, d]$ and $[o, b, b_2, d]$. As expected, the IIA property holds between paths within

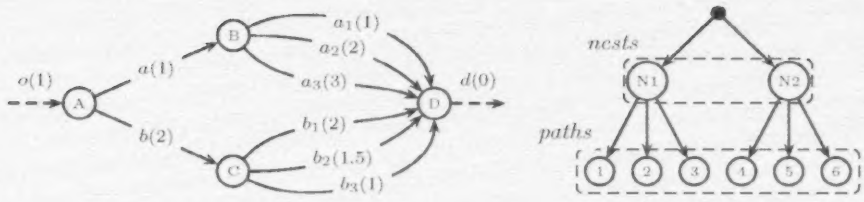

Figure 2: Illustrative network

Paths	Original	Probabilities with link removed			
		a_1	a_2	b_1	b_2
$1:[o,a,a_1,d]$	0.54	-	0.65(+20%)	0.55(+1%)	0.56(+4%)
$2:[o,a,a_2,d]$	0.15	0.38(+151%)	-	0.16(+1%)	0.16(+4%)
$3:[o,a,a_3,d]$	0.04	0.11(+151%)	0.05(+20%)	0.05(+1%)	0.05(+4%)
$4:[o,b,b_1,d]$	0.02	0.05(+93%)	0.03(+15%)	-	0.03(+19%)
$5:[o,b,b_2,d]$	0.06	0.12(+93%)	0.07(+15%)	0.17(+6%)	-
$6:[o,b,b_3,d]$	0.17	0.33(+93%)	0.20(+15%)	0.18(+6%)	0.21(+19%)

Table 1: Change in probability when link is removed

the same nest but not for paths in different nests. For example, when link a_1 is removed, path $[o,a,a_1,d]$ is also removed and the probabilities of the paths in the first nest rise by 151% while the paths in the second nest rise by 93%.

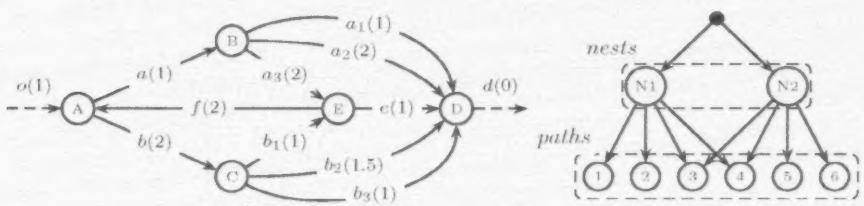

Figure 3: Illustrative network

The network in Figure 2 is designed so that the paths can naturally be divided into separate nests. In the next example shown in Figure 3 we slightly modify the network so that paths can be divided into a cross-nested structure (shown in the same figure). More precisely, we add a node E that separates links a_3 and b_1 into two links. The lengths of the considered paths in the new network do not change but the structure of the network is

different since apart from the origin and destination, two paths $[o, a, a_3, e, d]$ and $[o, a, b_1, e, d]$ share link e. There is a new link f going from node E to node A so that the expected maximum utilities from link a_3 and b_1 depend on the whole network and not only on the dummy link and link e.

We consider 6 main paths without loops: $[o, a, a_1, d]$, $[o, a, a_2, d]$, $[o, a, a_3, e, d]$, $[o, b, b_1, e, d]$, $[o, b, b_2, d]$, $[o, b, b_3, d]$, which are numbered as 1, 2, 3, 4, 5 and 6, respectively. We keep the same scales as in the first example (i.e. $\mu_a = 0.5$, $\mu_b = 0.8$ and the scale parameters are equal to one) in order to illustrate how the NRL model relaxes the IIA property when the network becomes more complicated. In Table 2 we report the changes in probabilities of the six paths when we remove in turn links a_3, b_1 and f. We note that the substitution patterns are different than in the previous example since the probabilities of paths 3 and 4 no longer change by the same proportion as the other paths in their respective nest.

Paths	Original	Probabilities with link removed		
		a_1	b_3	f
$1 : [o, a, a_1, d]$	0.54	-	0.60(+12%)	0.54(+0.7%)
$2 : [o, a, a_2, d]$	0.15	0.38(+150%)	0.17(+12%)	0.15(+0.7%)
$3 : [o, a, a_3, e, d]$	0.05	0.11(+148%)	0.05(+11%)	0.04(-1.3%)
$4 : [o, b, b_1, e, d]$	0.03	0.05(+86%)	0.05(+90%)	0.02(-6.7%)
$5 : [o, b, b_2, d]$	0.06	0.12(+93%)	0.12(+91%)	0.06(+1.4%)
$6 : [o, b, b_3, d]$	0.17	0.33(+93%)	-	0.17(+1.4%)

Table 2: Change in probability when link is removed

In order to compare the results with path based models we report probabilities given by the nested logit, cross-nested logit and link-nested logit (Vovsha and Bekhor, 1998) models in Table 3. For all models, the nesting parameters take the same values as in the NRL mode, namely 0.8 for nest $N1$ and 0.5 for nest $N2$. In the cross-nested logit model the inclusion coefficients α_{ij} define to which degree path i belong to nest j. We assign values so that the probabilities are as close as possibly to the NRL model ($\alpha_{31} = 1$, $\alpha_{32} = 0$, $\alpha_{41} = 0.4$, $\alpha_{42} = 0.6$). The cross-nested logit structure in Figure 3 is slightly different compared to the link-nested logit model, which is shown in Figure 4. The results show that for this example probabilities of the link-nested logit are slightly different from NRL. Moreover, the cross-nested logit probabilities can be very close to NRL for some values of the inclusion coefficients. Finally we note that the sum of the path probabilities for RL and NRL in the second example are numerically close but not exactly one. This is due to the cycle in the network.

In summary, the IIA property can be relaxed by assuming different scales. The resulting substitution pattern depends on the network structure but for

simple networks the model can be interpreted as a nested logit over paths.

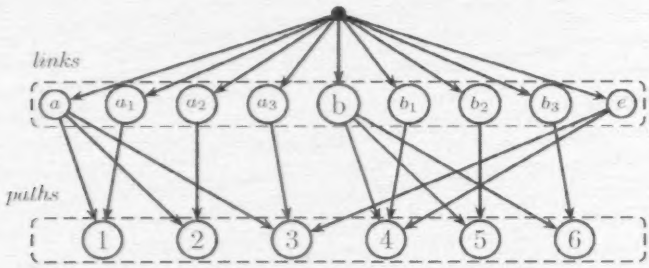

Figure 4: Cross-nested structure from the Link-nested logit model

Paths	Example 1			Example 2			
	MNL	NRL	Nested logit	MNL	NRL	Cross nested logit	Link nested logit
1	0.449	0.541	0.541	0.443	0.537	0.537	0.501
2	0.165	0.155	0.155	0.163	0.154	0.154	0.150
3	0.061	0.044	0.044	0.060	0.045	0.044	0.051
4	0.061	0.023	0.023	0.060	0.025	0.023	0.043
5	0.100	0.064	0.064	0.099	0.063	0.065	0.085
6	0.165	0.173	0.173	0.163	0.170	0.177	0.171

Table 3: Path probabilities comparison

4 Computation of the value functions

The main challenge associated with the NRL model is to efficiently solve system of system of non-linear equations (5). In the following we propose a solution method based on (i) value iterations with a good initial solution and (ii) dynamic accuracy.

We define a matrix $X(z)$ with entries

$$X(z)_{ka} = z_a^{\mu_a/\mu_k} \quad \forall k, a \in \tilde{A} \tag{10}$$

so that the Bellman equations (7) can be written as

$$z = [M \circ X(z)]e + b. \tag{11}$$

b is a vector of size $(|\tilde{A}| \times 1)$ with zero values for all states except for the destination that equals 1, e is a vector of size $(|\tilde{A}| \times 1)$ with value one for all states and $\circ$ is the element-by-element product.

Value iterations consist of solving Equation (11) repeatedly. We start with an initial vector z^0 and then for each iteration i we compute a new vector

$$z^{i+1} \leftarrow [M \circ X(z^i)]e + b. \tag{12}$$

and iterate until a fixed point is found using $||z^{i+1} - z^i||^2 < \gamma$ for a given threshold $\gamma > 0$ as stopping criteria. [1] It can be shown that if the Bellman equation has a solution, this method converges after a finite number of iterations. The choice of initial vector is however important for the rate of convergence. We use the solution of the system of linear equations corresponding to the RL model ($\mu_k = \mu \ \forall k \in A$) which is very fast to compute.

Since the value functions depend on the parameter values they need to be solved repeatedly when searching over the parameter space (maximum likelihood estimation). In order to speed up we use dynamic accuracy. More precisely, we update the threshold γ in the iterations of the non-linear optimization algorithm so that higher accuracy is required close to optimum (γ decreases as the number of iterations of the non-linear optimization algorithm increases). In the following section we discuss the maximum likelihood estimation in detail.

5 Maximum likelihood estimation

There are several different ways of estimating a dynamic discrete choice model (Aguirregabiria and Mira, 2010), we adopt the nested fixed point algorithm of Rust (1987). This algorithm combines an outer iterative non-linear optimization algorithm for searching over the parameter space with an inner algorithm for solving the value functions.[2] The latter was the focus of the previous section and we now turn our attention to the definition of the log-likelihood (LL) function and the derivation of its gradient which allows us to use classic Hessian approximation such as BHHH and BFGS (see for instance Berndt et al., 1974, Nocedal and Wright, 2006).

The path probabilities are defined by (8) and contain scale parameters $\mu_k \ \forall k \in A$ as well as the parameters β associated with the attributes of

[1] The value functions can also be used in the stopping criteria i.e. the iteration stops when $\sum_{k \in \tilde{A}} (V^{i+1}(k) - V^i(k))^2 < \gamma'$. The value functions have however larger magnitudes than z.

[2] An other option is the swapped nested fixed point algorithm of Aguirregabiria and Mira (2002). The idea is to swap the order of the outer and inner algorithms so that the outer algorithm solves the value functions and the inner algorithm maximizes the pseudo-likelihood function. This is very useful if the value functions are costly to evaluate which is not the case of the NRL model. Indeed, for NRL it is more costly to maximize the log-likelihood function than solving the value functions.

the instantaneous utilities. Clearly, it is not possible for a real network to estimate all link-specific scale parameters so we assume, without loss of generality, that they are a function of parameters β to be estimated $\mu_k(\beta)$. (We refer the reader to the numerical results, Section 6, for an example.) In the following we simplify the notation for the instantaneous utilities $v(k|a)$, value functions $V(k)$ and scale parameters μ_k (instead of $v(a|k;\beta)$, $V(k;\beta)$ and $\mu_k(\beta)$, respectively) but we emphasize that they depend on the vector of parameters β.

The LL function defined over the set of path observations $n = 1, \ldots, N$ is

$$LL(\beta) - \sum_{n=1}^{N} \ln P(\sigma_n, \beta) - \sum_{n=1}^{N} \sum_{t=0}^{I_n} \frac{1}{\mu_{k_t}} \left(v^n(k_{t+1}|k_t) + V^n(k_{t+1}) - V^n(k_t) \right) \tag{13}$$

and is very similar to the LL function of the RL model except that the value functions for each state do not cancel out. Assuming a linear-in-parameters formation of the instantaneous utilities, the gradient with respect to a given parameter β_i is

$$\frac{\partial LL(\beta)}{\beta_i} = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{I_n - 1} \frac{1}{\mu_{k_t}} \left(\frac{\partial v^n(k_{t+1}|k_t)}{\partial \beta_i} + \frac{\partial V^n(k_{t+1})}{\partial \beta_i} - \frac{\partial V^n(k_t)}{\partial \beta_i} \right) - \frac{\partial \mu_{k_t}}{\mu_{k_t}^2 \partial \beta_i} \left(v^n(k_{t+1}|k_t) + V^n(k_{t+1}) + V^n(k_t) \right)$$

and hence requires the first derivative of the value functions $V^n(k)$, $\forall k \in \tilde{A}$ with respect to β_i. We define $\phi_{ka} = \mu_a / \mu_k$ and take the derivative of a given value function z_k as defined by Equation (7) (without using the superscription for destination d) and obtain

$$\begin{aligned} \frac{\partial z_k}{\partial \beta_i} &= \sum_{a \in A} \left(\frac{\partial M_{ka}}{\partial \beta_i} z_a^{\phi_{ka}} + M_{ka} z_a^{\phi_{ka}} \left(\frac{\phi_{ka}}{z_a} \frac{\partial z_a}{\partial \beta_i} + \frac{\partial \phi_{ka}}{\partial \beta_i} \ln z_a \right) \right) \\ &= \sum_{a \in A} \left(\frac{\partial M_{ka}}{\partial \beta_i} z_a^{\phi_{ka}} + M_{ka} z_a^{\phi_{ka}} \frac{\partial \phi_{ka}}{\partial \beta_i} \ln z_a \right) + \sum_{a \in A} \left(M_{ka} z_a^{\phi_{ka}} \frac{\phi_{ka}}{z_a} \frac{\partial z_a}{\partial \beta_i} \right). \end{aligned} \tag{14}$$

We note that when the scales μ_k contains some model parameters, the derivative of each element of matrix $M(\beta)$ with respect to a given parameter β_i is

$$\frac{\partial M_{ka}}{\partial \beta_i} = \delta(a|k) e^{\frac{v(a|k)}{\mu_k}} \left(\frac{\partial v(a|k)}{\mu_k \partial \beta_i} - v(a|k) \frac{\partial \mu_k}{\mu_k^2 \partial \beta_i} \right), \quad k, a \in \tilde{A}$$

We introduce two matrices, G^i and K of size $|\tilde{A}| \times |\tilde{A}|$, which have the two sums of Equation (14) as entries

$$G^i_{ka} = \frac{\partial M_{ka}}{\partial \beta_i} z_a^{\phi_{ka}} + M_{ka} z_a^{\phi_{ka}} \frac{\partial \phi_{ka}}{\partial \beta_i} \ln z_a$$

$$K_{ka} = M_{ka} z_a^{\phi_{ka}} \frac{\phi_{ka}}{z_a}, \ \forall k, a \in \tilde{A}. \tag{15}$$

This allows us to define the Jacobian of vector z as a system of linear equations

$$\frac{\partial z}{\partial \beta_i} = G^i e + K \frac{\partial z}{\partial \beta_i} \Rightarrow \frac{\partial z}{\partial \beta_i} = (I - K)^{-1} G^i e. \tag{16}$$

which in theory can be solved very efficiently. In the other hand, it is possible to use the fact that $V(k) = \mu_k \ln z_k \ \forall k \in \tilde{A}$ and derive the Jacobian of V instead of z. In this case the gradient of $V(k)$ with respect to a given β_i is

$$\frac{\partial V(k)}{\partial \beta_i} = \frac{\partial \mu_k}{\partial \beta_i} \ln z_k + \frac{\mu_k}{z_k} \frac{\partial z_k}{\partial \beta_i}. \tag{17}$$

Using (14) we get

$$\frac{\partial V(k)}{\partial \beta_i} = \sum_{a \in A} S^i_{ka} + \sum_{a \in A} H_{ka} \frac{\partial V(a)}{\partial \beta_i} + h_k \tag{18}$$

where

$$S^i_{ka} = \mu_k \frac{\partial M_{ka}}{\partial \beta_i} \frac{z_a^{\phi_{ka}}}{z_k} + \mu_k M_{ka} \ln(z_a) \frac{z_a^{\phi_{ka}}}{z_k} \frac{\partial \phi_{ka}}{\partial \beta_i} - M_{ka} \ln(z_a) \frac{z_a^{\phi_{ka}}}{z_k} \frac{\partial \mu_a}{\partial \beta_i}$$

and

$$H_{ka} = M_{ka} \frac{z_a^{\phi_{ka}}}{z_k} \text{ and } h_k = \frac{\partial \mu_k}{\partial \beta_i} \ln z_k.$$

We denote S^i, H be two matrices of size $|\tilde{A}| \times |\tilde{A}|$ and h, V be two vectors of size $|\tilde{A}| \times 1$ with entries S^i_{ka}, H_{ka}, h_k, $V(k)$ for all $k, a \in \tilde{A}$, respectively. The Jacobian of vector V can then be written as a system of linear equations

$$\frac{\partial V}{\partial \beta_i} = (I - H)^{-1}(S^i e + h) \tag{19}$$

Theoretically, two formulas (16) and (19) can be used to compute the gradient of the value functions. To explain further the properties of these approaches we consdier the definitions of the matrix K and H in formulas (16) and (19). We note that z_a, $a \in \tilde{A}$ are exponential functions of the value

functions which are negative based on the assumption of the NRL model. The value of z_a may therefore be very close to zero. Since the elements of matrix K can be written as $K_{ka} = \phi_{ka} M_{ka} z_a^{\phi_{ka}-1}$ $(\forall k, a \in \tilde{A})$ if $\phi_{ka} < 1$, the value of K_{ka} can be very large and otherwise if $\phi_{ka} > 1$, K_{ka} can be very close to zero. The gaps between elements in matrix K (and also in matrix $I - K$) can lead to inaccurate solutions when solving the system (16). Based on equation (7), each element of matrix H can be written as

$$\begin{aligned} H_{ka} &= \frac{M_{ka} z_a^{\phi_{ka}}}{\sum_{a' \in A(k)} M_{ka'} z_{a'}^{\phi_{ka'}}} \\ &= \frac{1}{1 + \sum_{a' \in A(k), a' \neq a} \frac{M_{ka'} z_{a'}^{\phi_{ka'}}}{M_{ka} z_a^{\phi_{ka}}}} \\ &\forall k, a \in \tilde{A}, a \in A(k) \end{aligned}$$

so that $0 < H_{ka} < 1$, meaning that the gaps between elements of matrix H are smaller, compared to matrix K. This is helpful to avoid numerical issues when solving the system of linear equations. Therefore using formula (19) to compute the gradient of LL function is better than (16). In summary, the analytical gradient of the LL function has a complicated for but can be efficiently computed by solving systems of linear equations.

6 Numerical results

In this section we present estimation and prediction results for four different models: the RL model with and without link size (LS) attribute and the NRL model, also with and without LS attribute. We use the same data as Fosgerau et al. (2013) (also used in Frejinger and Bierlaire, 2007, Mai et al., 2014) which has been collected in Borlänge, Sweden. The network is composed of 3077 nodes and 7459 links and is uncongested so travel times can be assumed static and deterministic. The sample consists of 1832 trips corresponding to simple paths with a minimum of five links. Moreover, there are 466 destinations, 1420 different origin-destination (OD) pairs and more than 37,000 link choices in this sample.

6.1 Model specifications

The same five attributes as Fosgerau et al. (2013) are used in the instantaneous utilities. First, link travel time $TT(a)$ of action a. Second, a left turn dummy $LT(a|k)$ that equals one if the turn angle from k to a is larger than

40 degrees and less than 177 degrees. Third, a u-turn dummy $UT(a|k)$ that equals one if the turn angle is larger than 177. Fourth, a link constant $LC(a)$. The fifth attribute is $LS(a)$ (for a detailed description see Fosgerau et al., 2013) and it has been computed using a linear in parameters formulation of the aforementioned four attributes using parameters $\tilde{\beta}_{TT} = -2.5$, $\tilde{\beta}_{LT} = -1$, $\tilde{\beta}_{LC} = 0.4$, $\tilde{\beta}_{UT} = -4$.

There are over 7000 links in the network and it is hence not possible to estimate a link specific scale parameter. Instead, we define $\mu_k > 0$ to be an exponential function of link specific attributes. In order to make it easier to interpret the results, we use ω instead of β to denote the scale related parameters. The scales μ_k therefore can be written as $\mu_k(\omega) = e^{\lambda(\omega)}$. This assumption ensures that (i) the estimation problem is unconstrained and (ii) we can use the analytical gradient (17). Note that if all the parameters of the function $\lambda_k(.)$ are zero, the scales are equal to one for all links $k \in \tilde{A}$, meaning that the NRL model becomes the RL model.

For the numerical results presented in this paper we use the two link specific attributes available, travel time and LS, so $\lambda_k(\omega)$ is

$$\lambda_k(\omega) = \omega_{TT} TT(k) + \omega_{LS} LS(k). \tag{20}$$

We also note that we do not use link constant since it has the same value for all links. The rationale behind using it in the instantaneous utility is to penalize paths with many crossings (links).

To summarize, the deterministic utilities for four different model specifications with respect to link a given link k are

$$\begin{aligned} v^{\text{RL}}(a|k) = v^{\text{NRL}}(a|k;\beta) &= \beta_{TT} TT(a) + \beta_{LT} LT(a|k) + \beta_{LC} LC(a) \\ &\quad + \beta_{UT} UT(a|k) \\ v^{\text{RL LS}}(a|k) = v^{\text{NRL LS}}(a|k;\beta) &= \beta_{TT} TT(a) + \beta_{LT} LT(a|k) + \beta_{LC} LC(a) \\ &\quad + \beta_{UT} UT(a|k) + \beta_{LS} LS(a) \end{aligned}$$

and the instantaneous utilities are

$$\begin{aligned} u^{\text{RL}}(a|k;\beta) &= v^{\text{RL}}(a|k;\beta) + \mu\epsilon(a) \\ u^{\text{RL-LS}}(a|k;\beta) &= v^{\text{RL-LS}}(a|k;\beta) + \mu\epsilon(a) \\ u^{\text{NRL}}(a|k;\beta,\omega) &= v^{\text{NRL}}(a|k;\beta) + e^{\lambda_k(\omega)}\epsilon(a) \\ u^{\text{NRL-LS}}(a|k;\beta,\omega) &= v^{\text{NRL-LS}}(a|k;\beta) + e^{\lambda_k(\omega)}\epsilon(a). \end{aligned}$$

6.2 Estimation results

We report the estimation results for four specifications in Table 4. The β estimates have their expected signs and they are highly significant. Moreover,

their magnitudes are similar across the different models and the results are comparable to the ones previously published on the same data. $\widehat{\omega}_{TT}$ is not significantly different from zero while $\widehat{\omega}_{LS}$ is highly significant and negative. The LS attribute corresponds to expected normalized flows and takes positive values but is numerically close to zero for a majority of the links in the network. $\widehat{\omega}_{LS}$ shows that the scales are inversely related to flow so that links with more flow have larger variance of error terms than links with less flow.

There is remarkable improvement in final log-likelihood values when adding the LS attribute, which is also pointed in previous work (Fosgerau et al., 2013, Mai et al., 2014). The best model in terms of fit is NRL-LS.

Before comparing prediction results in the following section we make some remarks concerning the estimation. We use a basic trust region algorithm with the BHHH method for approximating the Hessian and the code is implemented in MATLAB (and available upon request). We use the iterative method with dynamic accuracy for the computation of the value functions (see Section 4). We note that if we use an initial vector as a solution of the system of linear equations, circa 100 iterations is enough for a high precision ($\gamma' = 10^{-8}$) but we need circa 200 iterations for the same precision when the initial vector is the unit vector (all the elements are equal to one). Moreover, using only 50 iterations in the beginning of the optimization (corresponding to a precision $\gamma' \in [1, 10]$) and switching to the high precision $\gamma' = 10^{-8}$ when the norm of the gradient of the log-likelihood function is less than 10^{-3} we observe a speed up of two times can be achieved.

6.3 Prediction results

In this section we focus on comparing the prediction performance of the different models. In this context, the predicted log-likelihood (PLL) values for holdout samples can be used as a performance measure. We use a cross validation approach where the sample of observations is divided into two sets. The set is used for estimation and consists of 80% of the path observations which are drawn randomly (uniform distribution). The second set (20% of the observations) is the holdout and the estimated model is used to evaluate the corresponding probabilities. We generate 40 holdout samples of the same size by reshuffling the real sample. More precisely, for each holdout sample i, $0 \leq i \leq 40$ we estimate the parameters $\hat{\beta}_i$ based on the training sample and this vector of parameters is used to compute the PLL value (PLL_i) for the prediction sample. So basically PLL_i is conditional on the holdout sample i. In order to have unconditional PLL values we compute the average of the

Parameters	RL	NRL	RL-LS	NRL-LS
$\widehat{\beta}_{TT}$	-2.494	-2.572	-3.060	-3.008
Rob. Std. Err.	0.098	0.099	0.103	0.103
Rob. t-test(0)	-25.45	-25.98	- 27.709	-29.204
$\widehat{\beta}_{LT}$	-0.933	-0.904	-1.057	-1.014
Rob. Std. Err.	0.030	0.030	-0.029	0.031
Rob. t-test(0)	-31.10	-30.13	-36.448	-32.710
$\widehat{\beta}_{LC}$	-0.411	-0.344	-0.353	-0.305
Rob. Std. Err.	0.013	0.014	0.011	0.013
Rob. t-test(0)	-31.62	-24.57	-32.091	-23.461
$\widehat{\beta}_{UT}$	-4.459	-4.442	-4.531	-4.468
Rob. Std. Err.	0.114	0.133	0.126	0.144
Rob. t-test(0)	-39.11	-33.40	- 35.960	-32.028
$\widehat{\beta}_{LS}$	-	-	-0.227	-0.212
Rob. Std. Err.	-	-	-0.013	0.013
Rob. t-test(0)	-	-	-17.462	-16.308
$\widehat{\omega}_{TT}$	-	0.307	-	0.114
Rob. Std. Err.	-	0.276	-	0.311
Rob. t-test(0)	-	1.11	-	0.367
$\widehat{\omega}_{LS}$	-	-0.946	-	-0.856
Rob. Std. Err.	-	0.088	-	0.087
Rob. t-test(0)	-	-10.75	-	-9.839
$LL(\widehat{\beta})$	-6303.9	-6212.3	-6045.6	-5976.0

Table 4: Estimation results

PLL values over samples as follows

$$\overline{PLL}_p = \frac{1}{p}\sum_{i=1}^{p} PLL_i \quad \forall 1 \leq p \leq 40 \tag{21}$$

The values of $\overline{PLL}_p$, $1 \leq p \leq 40$ are plotted in Figure 5 and Table (5) reports the average of the PLL values over 40 samples given by the RL, RL-LS, NRL, NRL-LS models. For each model the value of $\overline{PLL}_p$ become more stable as p increases. The prediction results show that models including LS perform better than those without. Even though the model fit is significantly better for NRL-LS than RL-LS the prediction results are similar.

Readers can refer to the Appendix for the details of the average of the PLL values estimated by 40 holdout samples.

RL	NRL	RL-LS	NRL-LS
-1241.54	-1240.00	-1190.36	-1189.52

Table 5: Average of PLL values over 40 holdout samples

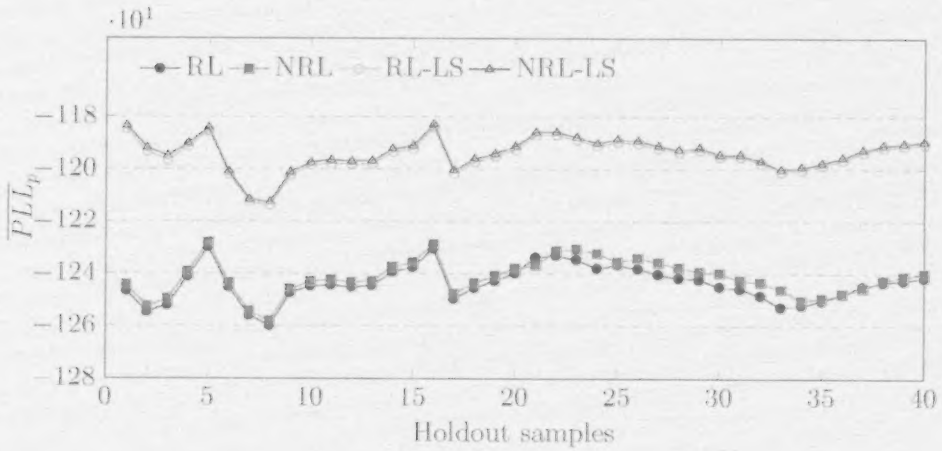

Figure 5: Average of the PLL values over holdout samples

7 Conclusion

This paper presents the NRL model that relaxes the IIA assumption of the RL model by allowing scale parameters to be link specific while keeping the advantages of the RL model. We propose an efficient way for estimating the model solving the value functions using an iterative method with dynamic accuracy. Moreover, we derive the gradient of the log-likelihood function which can be computed by solving systems of linear equations.

We provide numerical results using real data. The parameter estimates are sensible and the NRL model has remarkably better fit than the RL model. The LS attribute plays an important role and the best model in terms of fit is NRL combined with a LS attribute (NRL-LS). We provide a cross-validation study that shows that NRL-LS and RL-LS are the best models for prediction and their performance is very similar, unlike the model fit.

In future research we plan to further investigate further the importance of the LS attribute and its definition with a sensitivity analysis. Moreover, there are quite few attributes available in the data set used in this paper. We would like to test the model on other data sets to further study possible functional forms of the scale parameters.

Acknowledgement

This research was partly funded by the National Sciences and Engineering Research Council of Canada, discovery grant 435678-2013.

References

Aguirregabiria, V. and Mira, P. Swapping the nested fixed point algorithm: A class of estimators for discrete markov decision models. *Econometrica*, 70(4):1519–1543, 2002.

Aguirregabiria, V. and Mira, P. Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1):38–67, 2010.

Bekhor, S., Ben-Akiva, M., and Ramming, M. Estimating route choice models for large urban networks. 9th World Conference on Transport Research, Seoul, Korea, 2001.

Ben-Akiva, M. *The structure of travel demand models.* PhD thesis, MIT, 1973.

Berndt, E. K., Hall, B. H., Hall, R. E., and Hausman, J. A. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, 3/4:653–665, 1974.

Chu, C. A paired combinatorial logit model for travel demand analysis. In *Proceedings of the fifth World Conference on Transportation Research*, volume 4, pages 295–309, Ventura, CA, 1989.

Fosgerau, M., Frejinger, E., and Karlström, A. A link based network route choice model with unrestricted choice set. *Transportation Research Part B*, 56:70–80, 2013.

Frejinger, E., Bierlaire, M., and Ben-Akiva, M. Sampling of alternatives for route choice modeling. *Transportation Research Part B*, 43(10):984–994, 2009.

Frejinger, E. and Bierlaire, M. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B*, 41(3):363–378, 2007.

Guevara, C. A. and Ben-Akiva, M. E. Sampling of alternatives in multivariate extreme value (MEV) models. *Transportation Research Part B*, 48(1):31 – 52, 2013. ISSN 0191-2615.

Mai, T., Frejinger, E., and Bastin, F. A misspecification test for logit based route choice models. *Technical report, CIRRELT - 32*, 2014.

Nocedal, J. and Wright, S. J. *Numerical Optimization*. Springer, New York, NY, USA, 2nd edition, 2006.

Prato, C. G. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2:65–100, 2009.

Rust, J. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5):999–1033, 1987.

Train, K. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003.

Vovsha, P. and Bekhor, S. Link-nested logit model of route choice Overcoming route overlapping problem. *Transportation Research Record*, 1645: 133–142, 1998.

Appendix

Sample	$\overline{PLL_p}$			
	RL	NRL	RL-LS	NRL-LS
1	-1246.56	-1244.41	-1184.87	-1183.56
2	-1254.74	-1252.56	-1193.44	-1192.06
3	-1252.09	-1249.66	-1196.61	-1195.05
4	-1241.10	-1239.09	-1191.43	-1190.20
5	-1229.88	-1228.13	-1185.43	-1184.34
6	-1245.17	-1243.63	-1202.04	-1201.05
7	-1255.99	-1254.46	-1212.61	-1211.68
8	-1259.84	-1258.11	-1213.70	-1212.63
9	-1247.43	-1245.77	-1202.31	-1201.25
10	-1244.75	-1242.94	-1198.56	-1197.41
11	-1244.36	-1242.39	-1197.83	-1196.61
12	-1245.07	-1243.24	-1198.18	-1197.06
13	-1244.53	-1242.76	-1197.99	-1196.88
14	-1239.35	-1237.50	-1193.45	-1192.27
15	-1237.55	-1235.67	-1192.25	-1191.04
16	-1230.43	-1228.66	-1183.98	-1182.90
17	-1249.51	-1247.74	-1201.62	-1200.55
18	-1245.44	-1243.57	-1197.24	-1196.13
19	-1242.70	-1240.87	-1195.23	-1194.13
20	-1239.73	-1237.90	-1192.48	-1191.34
21	-1233.83	-1236.78	-1186.99	-1185.96
22	-1233.02	-1231.30	-1186.86	-1185.91
23	-1234.65	-1230.69	-1188.85	-1187.85
24	-1237.97	-1232.21	-1191.12	-1190.11
25	-1236.66	-1235.40	-1189.67	-1188.63
26	-1238.11	-1234.13	-1190.16	-1189.25
27	-1240.30	-1235.68	-1192.12	-1191.16
28	-1241.72	-1237.77	-1193.54	-1192.64
29	-1242.21	-1239.26	-1192.78	-1191.92
30	-1245.08	-1239.78	-1195.39	-1194.54
31	-1245.85	-1242.59	-1195.33	-1194.55
32	-1248.45	-1243.44	-1197.78	-1196.97
33	-1252.83	-1245.97	-1200.97	-1200.14
34	-1252.09	-1250.20	-1200.15	-1199.30
35	-1250.31	-1249.49	-1198.63	-1197.78
36	-1247.90	-1247.82	-1196.45	-1195.59
37	-1244.89	-1245.48	-1193.53	-1192.70
38	-1243.23	-1242.61	-1191.71	-1190.91
39	-1242.58	-1241.02	-1191.10	-1190.26
40	-1241.54	-1240.00	-1190.36	-1189.52

Table 6: Prediction results